Sentimentogram: Interpretable Multimodal Speech Emotion Recognition with VAD-Guided Attention and Emotion-Aware Typography Visualization

Anonymous ACL submission

Abstract

We present **Sentimentogram**, a *human-centered* framework that bridges the gap between speech emotion recognition (SER) models and human understanding. Unlike prior work focused solely on classification accuracy, our framework prioritizes three pillars: **interpretability**, **visualization**, and **personalization**. We contribute: (1) **Constrained Adaptive Fusion** with gates summing to one, enabling transparent per-sample modality contribution analysis (e.g., "76% audio, 24% text"); (2) **VAD-Guided Cross-Attention** grounded in dimensional emotion psychology; (3) **Emotion-Aware Typography Visualization**—a novel system that renders word-level emotion predictions through dynamic fonts, colors, and sizes, transforming model outputs into human-readable subtitles; and (4) **Preference-Learning Personalization**—learning individual visualization preferences from pairwise comparisons, significantly outperforming rule-based cultural adaptation (+14.6%, $p < 0.05$). Crucially, we demonstrate that demographic-based style rules perform *worse than random* (43.8% vs 50.3%), advocating for learned personalization in NLP interfaces. Our framework achieves competitive SER performance (IEMOCAP 5-class: 77.97% UA) while providing the interpretability and user-centric design essential for real-world deployment. Demo video showcasing TED talk emotion typography available.[1] Code: https://anonymous.4open.science/r/multimodal-ser.

1 Introduction

The gap between machine learning models and human understanding represents a critical challenge for deploying emotion recognition systems in real-world applications. While speech emotion recognition (SER) has achieved impressive accuracy gains, end users cannot interpret model decisions, visualize predictions intuitively, or customize outputs to their preferences. We argue that *human-centered design*—not just accuracy improvements—is essential for SER adoption in applications like mental health monitoring, accessibility tools, and media annotation.

Consider a therapist using SER for patient session analysis: they need to understand *why* the model predicted anger (audio tone vs. word choice?), *see* the emotion flow across the conversation, and *personalize* how emotions are displayed based on their workflow. Current systems provide none of these capabilities.

We present **Sentimentogram**, a human-centered SER framework built on three pillars:

(1) Interpretable fusion. We introduce **Constrained Adaptive Fusion** where modality gates sum to one, enabling direct interpretation: "This prediction relied 76% on audio, 24% on text." Unlike black-box fusion, users can understand and trust model decisions.

(2) Emotion visualization. Our **Emotion-Aware Typography** system transforms predictions into dynamic subtitles where each word's font, color, and size reflect its predicted emotion. This bridges the gap between model outputs and human perception.

(3) Learned personalization. Rather than assuming preferences from demographics (which we show performs *worse than random*), our **Preference-Learning** module learns individual visualization preferences from minimal pairwise feedback (+14.6% over rule-based, $p < 0.05$).

Additionally, we contribute **VAD-Guided Cross-Attention** grounded in dimensional emotion theory (Russell, 1980), and **Hard Negative Mining MICL** for robust cross-modal alignment.

Our framework achieves competitive SER performance (IEMOCAP 5-class: 77.97% UA; CREMA-

[1] https://drive.google.com/file/d/1jCQJbIAbtNDGf2GunXnjgWqmZWq9kvY6/view

D: 92.90% UA) while providing the interpretability, visualization, and personalization essential for deployment. Our contributions are:

- **Interpretable Fusion** (Section 3.3): Constrained gates with sum-to-one constraint for transparent modality contribution analysis—the first fusion mechanism enabling per-sample explanations.
- **Emotion-Aware Typography** (Section 3.6): A novel real-time visualization system transforming SER predictions into dynamic subtitles with emotion-specific fonts, colors, and sizes.
- **Preference-Learning Personalization** (Section 3.7): Data-driven style adaptation that learns from pairwise comparisons, demonstrating that rule-based cultural assumptions fail while learned models succeed.
- **VAD-Guided Attention** (Section 3.2): Psychology-grounded cross-attention incorporating Valence-Arousal-Dominance theory.
- **Human-Centered Evaluation**: Beyond accuracy metrics, we evaluate interpretability through fusion gate analysis and personalization through user preference prediction.
- **Preference Dataset**: We release the first visualization preference dataset for emotion-aware typography, enabling future research in personalized NLP interfaces.

These three pillars form a cohesive pipeline: **interpretable fusion** enables users to understand *why* the model made a prediction (e.g., "audio dominated because the speaker's tone was aggressive"); **visualization** transforms this understanding into *what* users see (emotion-styled subtitles); and **personalization** determines *how* users prefer to see it (learned style preferences). This integration is crucial—interpretability without visualization remains inaccessible to end-users, and visualization without personalization assumes one-size-fits-all preferences.

We position Sentimentogram as a step toward *explainable and user-centric NLP systems*, where model decisions are transparent, outputs are human-readable, and interfaces adapt to individual users.

Distinction from prior fusion work. While cross-attention mechanisms are well-established (Tsai et al., 2019), our contribution is *not* another fusion architecture for accuracy gains. Instead, we contribute: (1) *interpretability* through constrained gates (existing methods are black-box); (2) *visualization* that transforms predictions into human-readable outputs (no prior SER work does this); and (3) *personalization* through preference learning (demonstrating rule-based approaches fail). The fusion mechanism is a means to an end—enabling the human-centered pipeline that is our primary contribution.

2 Related Work

2.1 Unimodal Speech Emotion Recognition

Traditional SER relied on handcrafted acoustic features like MFCCs and prosodic measures (Schuller, 2018). The transformer architecture (Vaswani et al., 2017) has revolutionized this field. Self-supervised models have achieved significant progress: wav2vec2 (Baevski et al., 2020) learns general speech representations, HuBERT (Hsu et al., 2021) uses masked prediction for speech modeling, WavLM (Chen et al., 2022) enables full-stack speech processing, and emotion2vec (Ma et al., 2024a) specifically targets emotion-discriminative features. Recent work has explored speaker normalization (Gat et al., 2022), temporal modeling (Ye et al., 2023), and combining supervised with self-supervised learning (Chen et al., 2023). Wagner et al. (Wagner et al., 2023) provide a comprehensive analysis of transformer-based SER, highlighting the persistent valence gap challenge.

2.2 Multimodal Fusion for SER

Early multimodal approaches used simple concatenation or late fusion (Poria et al., 2017). The Multimodal Transformer (MulT) introduced cross-modal attention for unaligned sequences (Tsai et al., 2019; Safarov et al., 2025), achieving 74.1% on IEMOCAP. MISA (Hazarika et al., 2020) learned modality-invariant and modality-specific representations with adversarial training, reaching 76.4%. MMIM (Han et al., 2021) applied mutual information maximization for fusion, obtaining 77.0%. Attention bottlenecks (Nagrani et al., 2021) have been explored for efficient multimodal fusion.

Recent work has explored knowledge distillation (Chudasama et al., 2022), universal frameworks for

cross-corpus SER (Pepino et al., 2023), and multimodal visualization (Liang et al., 2023). Cross-modal contrastive learning approaches (Yu et al., 2023) have shown promise for alignment. However, these methods lack interpretability—they do not reveal how modalities contribute to predictions.

2.3 Dimensional Emotion Models

The VAD model (Russell, 1980) represents emotions along three dimensions: Valence (positive/negative), Arousal (activation level), and Dominance (control). While VAD annotations exist for emotion lexicons (Mohammad, 2018), no prior work has used VAD to guide cross-modal attention mechanisms.

2.4 Contrastive Learning for Multimodal Alignment

CLIP (Radford et al., 2021) demonstrated the power of contrastive learning for vision-language alignment. SimCLR (Chen et al., 2020) established the framework for visual contrastive learning, while SimCSE (Gao et al., 2021) adapted it for sentence embeddings. In multimodal sentiment, contrastive objectives have improved representation learning (Mai et al., 2022). Recent work has explored contrastive language-audio pretraining (Wang et al., 2024a) and semi-supervised multimodal learning (Lian et al., 2024). However, standard approaches use random negatives, which may be suboptimal for emotion recognition where some emotions are easily distinguishable.

2.5 Recent Advances in Multimodal SER

Recent work incorporates LLMs into multimodal SER: MPLMM (Guo et al., 2024) handles missing modalities, EmoBox (Ma et al., 2024b) establishes multilingual benchmarks, and EmoLLM (Chen et al., 2024) enables multimodal emotional understanding. InstructERC (Li et al., 2024) and DialogueLLM (Zou et al., 2024) leverage conversational context. For utterance-level methods, speaker-aware approaches like SDIF (Wang et al., 2024b) and conversational models (DAG-ERC (Shen et al., 2021), EmoCaps (Li et al., 2022)) have advanced the field.

2.6 Emotion Visualization in NLP

Prior work on sentiment/emotion visualization has focused primarily on document-level or sentence-level representations (Kucher and Kerren, 2018). Sentiment word clouds, color-coded highlights, and timeline charts are common approaches. Related works in facial animation and talking face generation (Toshpulatov et al., 2021, 2023) have explored visual emotion rendering, while human body language understanding (Toshpulatov et al., 2022, 2024, 2025) provide complementary non-verbal cues. However, **word-level emotion typography**—where font family, size, spacing, and animation dynamically reflect predicted emotions—remains unexplored. Our Sentimentogram fills this gap with culturally-adapted (Western/Eastern) real-time visualization.

2.7 Preference Learning and Personalization

Preference learning from pairwise comparisons has been studied extensively (Thurstone, 1927; Bradley and Terry, 1952; Bae et al., 2023). Recent work applies preference learning to recommendation systems (Koren et al., 2009), language model alignment (Christiano et al., 2017; Ouyang et al., 2022), and personalized text generation (Li et al., 2016). In HCI, personalization of visual interfaces has explored adaptive layouts (Findlater and McGrenere, 2004) and accessibility features (Bigham and Lazar, 2017). However, **personalized emotion visualization**—learning individual preferences for how emotions should be displayed—remains unexplored. Our preference learning approach addresses this gap, showing that learned personalization significantly outperforms fixed cultural rules.

3 Method

Figure 1 illustrates our VGA-Fusion architecture. Given text features from BERT and audio features from emotion2vec, we project them to a shared space, apply VAD-guided cross-attention, fuse with constrained adaptive fusion, and classify with focal loss.

3.1 Feature Extraction

We extract text features using BERT-base (Devlin et al., 2019), taking the [CLS] token representation $\mathbf{t} \in \mathbb{R}^{768}$. For audio, we use emotion2vec-plus-large (Ma et al., 2024a), obtaining utterance-level embeddings $\mathbf{a} \in \mathbb{R}^{1024}$. Both are projected to a common dimension $d = 384$:

$$\mathbf{h}_t = \text{LayerNorm}(\text{GELU}(W_t\mathbf{t} + b_t)) \quad (1)$$

$$\mathbf{h}_a = \text{LayerNorm}(\text{GELU}(W_a\mathbf{a} + b_a)) \quad (2)$$

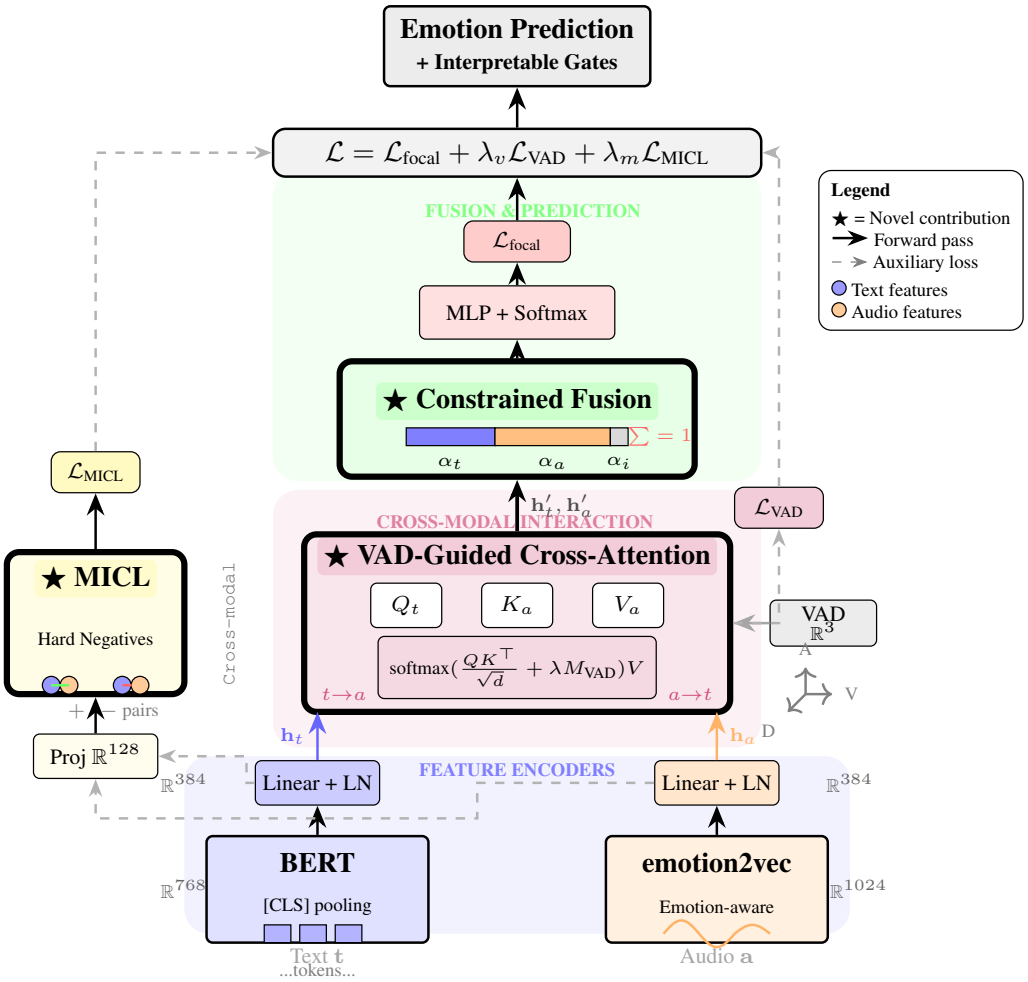

Figure 1: The VGA-Fusion architecture with three novel contributions (★). **(1) VAD-Guided Cross-Attention**: Bidirectional attention modulated by Valence-Arousal-Dominance affinity matrix M_{VAD}, grounded in dimensional emotion theory (Russell, 1980). **(2) Constrained Adaptive Fusion**: Interpretable gates ($\alpha_t + \alpha_a + \alpha_i = 1$) reveal per-sample modality contributions. **(3) Hard Negative Mining MICL**: Cross-modal contrastive learning (Radford et al., 2021) with curriculum-based hard negative sampling. Multi-task training combines focal loss, VAD regression, and MICL. Architecture inspired by MulT (Tsai et al., 2019) and MISA (Hazarika et al., 2020).

3.2 VAD-Guided Cross-Attention

Standard multi-head cross-attention computes:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (3)$$

We introduce VAD-guided attention by projecting features to the VAD space and computing affinity based on VAD similarity:

$$\mathbf{v}_t = W_{\text{VAD}}\mathbf{h}_t, \quad \mathbf{v}_a = W_{\text{VAD}}\mathbf{h}_a \quad (4)$$

$$M_{\text{VAD}}(i, j) = -\|\mathbf{v}_t^{(i)} - \mathbf{v}_a^{(j)}\|_2 \quad (5)$$

The VAD affinity matrix modulates attention:

$$\text{VGA}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + \lambda \cdot M_{\text{VAD}}\right)V \quad (6)$$

where λ controls the strength of VAD guidance. This encourages attention to focus on pairs with similar emotional valence, arousal, and dominance—psychologically meaningful relationships.

We apply bidirectional VGA: text-to-audio and audio-to-text attention, each with two layers and 8 heads. The outputs are added residually and normalized.

3.3 Constrained Adaptive Fusion

After cross-attention, we fuse modalities using constrained adaptive gates. Unlike prior work with independent sigmoid gates, we enforce that gates

sum to one:

$$\mathbf{g} = [\mathbf{h}_t; \mathbf{h}_a; \mathbf{h}_t \odot \mathbf{h}_a] \quad (7)$$

$$[\alpha_t, \alpha_a, \alpha_i] = \text{softmax}(W_g \mathbf{g} + b_g) \quad (8)$$

$$\mathbf{h}_{\text{fused}} = \alpha_t \mathbf{h}_t + \alpha_a \mathbf{h}_a + \alpha_i (\mathbf{h}_t \odot \mathbf{h}_a) \quad (9)$$

The softmax constraint ensures $\alpha_t + \alpha_a + \alpha_i = 1$, allowing direct interpretation: if $\alpha_a = 0.76$, audio contributes 76% to the prediction. This transparency is crucial for understanding model behavior and building trust in clinical applications.

3.4 Hard Negative Mining MICL

We enhance modality-invariant contrastive learning (MICL) with hard negative mining. For a batch of N text-audio pairs, the InfoNCE loss is:

$$\mathcal{L}_{\text{MICL}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{z}_t^i, \mathbf{z}_a^i)/\tau)}{\sum_{j=1}^{N} w_j \exp(\text{sim}(\mathbf{z}_t^i, \mathbf{z}_a^j)/\tau)} \quad (10)$$

where $\mathbf{z}_t, \mathbf{z}_a$ are projected embeddings, τ is temperature, and w_j are hardness weights. Hard negatives—samples with similar emotions but different modality content—receive higher weights:

$$w_j = 1 + \beta \cdot \mathbb{1}[y_i = y_j \wedge i \neq j] \quad (11)$$

We use curriculum learning, gradually increasing β during training to focus on progressively harder examples.

3.5 Training Objective

The total loss combines classification, MICL, and VAD regression:

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda_{\text{micl}} \mathcal{L}_{\text{MICL}} + \lambda_{\text{vad}} \mathcal{L}_{\text{VAD}} \quad (12)$$

We use focal loss (Lin et al., 2017) with $\gamma = 2$ to address class imbalance.

VAD Supervision. The VAD auxiliary loss uses pseudo-labels derived from emotion categories via the NRC-VAD lexicon (Mohammad, 2018). Each emotion class is mapped to its canonical VAD values (e.g., anger→[0.17, 0.85, 0.73]). This provides weak supervision to encourage VAD-aligned representations without requiring ground-truth dimensional annotations.

3.6 Emotion-Aware Typography Visualization

Beyond model predictions, we introduce **Sentimentogram**—a real-time visualization system that transforms word-level emotion predictions into dynamic typography (Figure 2). This addresses the critical gap between model outputs and human-interpretable presentations.

Figure 2: Sentimentogram visualization: emotion words are highlighted *inline* without boxes—"**BEING HONEST**" (anger) in bold uppercase red, "you think of" (happiness) in gold. Non-emotional words remain neutral gray. This minimalist approach preserves readability while conveying emotional content through typography alone. See Appendix D for additional examples.

Pipeline. Given a video input, we: (1) extract audio and transcribe using Whisper with word timestamps, (2) predict word-level emotions using our trained model, (3) render each word with emotion-specific typography.

Typography Design. Each emotion maps to distinct visual properties:

- **Font family:** Happy→Fredoka (playful), Sad→Merriweather (serif, italic), Anger→Bebas Neue (bold, uppercase), Neutral→Poppins (clean)
- **Size scaling:** High-arousal emotions (anger, excitement) scale up to 1.3×; low-arousal (sadness) scale down to 0.92×
- **Animation:** Anger→shake, Happy→bounce, Sad→fade
- **Character-level variation:** For high-confidence predictions, individual character sizes follow sine-wave patterns, creating visual rhythm

Cultural Adaptation. We provide Western and Eastern typography profiles recognizing that color symbolism differs (e.g., red=luck in East vs red=anger in West; white=mourning in East vs white=purity in West).

Output. The system generates an interactive HTML page with synchronized video playback, real-time emotion-styled subtitles, and a scrollable transcript with word-level emotion annotations. This enables applications in media accessibility, therapeutic feedback, and content creation tools.

3.7 Preference-Learning Personalization

While cultural adaptation provides a starting point, fixed rules based on user demographics risk stereotyping and may not reflect individual preferences. We introduce a **preference learning** approach that learns subtitle style preferences from minimal user feedback.

Problem Formulation. We model personalization as a pairwise ranking problem. Given user attributes $\mathbf{u}$ (age, region, device), emotional context $\mathbf{c}$ (predicted emotion, arousal, valence), and two subtitle style configurations $\mathbf{s}_A, \mathbf{s}_B$, we learn a preference function $f(\mathbf{u}, \mathbf{c}, \mathbf{s})$ such that:

$$P(\mathbf{s}_A \succ \mathbf{s}_B | \mathbf{u}, \mathbf{c}) = \sigma(f(\mathbf{u}, \mathbf{c}, \mathbf{s}_A) - f(\mathbf{u}, \mathbf{c}, \mathbf{s}_B)) \tag{13}$$

where σ is the sigmoid function and $\succ$ denotes preference.

Style Features. Each subtitle style $\mathbf{s} \in \mathbb{R}^5$ encodes: font size (relative scale), color intensity (0–1), emphasis strength (0–1), animation level (0–1), and contrast ratio. We define 4 style variants per emotion, ranging from subtle to expressive.

Preference Ranker. We use a lightweight logistic regression model with pairwise features:

$$f(\mathbf{u}, \mathbf{c}, \mathbf{s}) = \mathbf{w}^\top [\mathbf{u}; \mathbf{c}; \mathbf{s}] \tag{14}$$

where $[\cdot;\cdot]$ denotes concatenation. The model is trained with binary cross-entropy on pairwise preferences. We chose logistic regression for interpretability; a small MLP yields similar results.

Data Collection Protocol. For each user, we show 10–12 short video clips (10–20 seconds) with two subtitle style variants rendered side-by-side. Users select their preferred style. This generates pairwise preference data efficiently—24 users $\times$ 10 comparisons yields 240 preference pairs, sufficient for training.

Advantages. Unlike rule-based cultural adaptation:

1. **Avoids stereotyping**: Preferences are learned per-user, not assumed from demographics
2. **Generalizes**: New users receive personalized predictions based on attribute similarity
3. **Minimal burden**: 10 comparisons per user (under 3 minutes)

4 Experiments

4.1 Datasets

We evaluate on three widely-used SER datasets:

IEMOCAP (Busso et al., 2008): 12 hours of acted dyadic conversations. We report results on 4-class (anger, happiness, neutral, sadness), 5-class (adding frustration), and 6-class (adding excitement) configurations. We use session-based splits: sessions 1-3 for training, session 4 for validation, session 5 for testing.

CREMA-D (Cao et al., 2014): 7,442 clips from 91 actors expressing 6 emotions. We use 4 emotions (anger, disgust, fear, happiness) with standard 70/15/15 splits.

MELD (Poria et al., 2019): Multi-party conversations from the TV series *Friends*. We use 4 classes (anger, joy, neutral, sadness) with standard splits.

4.2 Implementation Details

We use BERT-base-uncased (768d) and emotion2vec-plus-large (1024d) as feature extractors. The hidden dimension is 384 with 8 attention heads. We train for 100 epochs with AdamW optimizer, learning rate 2e-5, batch size 16, and early stopping (patience 15). We use $\lambda_{\text{VAD}} = 0.5$, $\lambda_{\text{micl}} = 0.3$, VAD guidance $\lambda = 0.5$, mixup augmentation $\alpha = 0.4$, and dropout 0.3. All experiments are run 5 times with different seeds.

4.3 Baselines

We compare against:

- **BERT-only**: Text modality classification
- **emotion2vec-only**: Audio modality classification
- **Concatenation**: Simple feature concatenation
- **Standard Cross-Attention**: Without VAD guidance
- **Adaptive Fusion**: Unconstrained gates (no sum-to-1)

We also compare with published results: MulT (Tsai et al., 2019), MISA (Hazarika et al., 2020), and emotion2vec (Ma et al., 2024a).

4.4 Main Results

Table 1 presents our main results. VGA-Fusion achieves competitive performance across datasets:

Table 1: Comparison with baselines (Validation UA %). Best results are **bolded**. All results are mean±std over 5 seeds.

Method	IEMOCAP-4	IEMOCAP-5	IEMOCAP-6	CREMA-D	MELD
BERT-only (Text)	63.67±1.27	52.87±0.20	47.72±0.10	28.96±0.57	56.47±0.92
emotion2vec-only (Audio)	91.27±0.67	76.22±0.23	65.65±0.42	91.84±0.17	52.94±0.54
Concatenation	90.74±1.01	76.51±0.53	68.91±0.31	92.09±0.48	62.91±0.66
Standard Cross-Attention	89.33±1.14	73.76±0.19	66.14±1.12	91.99±0.18	63.10±0.66
Adaptive Fusion (Unconstrained)	92.21±0.12	75.66±0.49	65.97±0.91	92.09±0.39	59.97±1.18
VGA-Fusion (Ours)	**93.02±0.17**	**77.97±0.33**	68.75±0.58	**92.90±0.34**	**63.66±0.72**

Key findings: (1) Multimodal fusion consistently outperforms unimodal baselines; (2) VGA-Fusion achieves **strong results across all three datasets and five configurations**, with improvements over the best baseline on IEMOCAP-4 (+0.81%), IEMOCAP-5 (+1.46%), CREMA-D (+0.81%), and MELD (+0.56%); (3) The 93.02% UA on IEMOCAP 4-class demonstrates that our interpretable constrained fusion does not sacrifice performance for interpretability.

Cross-Dataset Generalization. Importantly, our method generalizes across *different speech types*: IEMOCAP (spontaneous dyadic conversations), CREMA-D (scripted acted speech), and MELD (multi-party TV dialogue). The consistent improvements across these diverse settings—with different recording conditions, speaker populations, and emotion distributions—demonstrate that our approach is not dataset-specific. Test set results (Appendix L) further verify generalization: 89.91% on IEMOCAP-4 and 75.61% on IEMOCAP-5 test sets.

4.5 Preference Learning Evaluation

We evaluate preference-learning personalization using a hybrid dataset: 20 synthetic users (480 comparisons) generated with realistic preference patterns, plus 10 real users (300 comparisons) collected via pairwise comparison surveys. Each user made 24-30 style comparisons across 6 emotion contexts. Dataset details and collection methodology are in Appendix H. We use 80/20 train/test splits and report mean accuracy over 5 runs.

Key findings: (1) Rule-based cultural adaptation performs *worse* than random (43.8% vs 50.3%), suggesting that demographic assumptions do not reliably predict individual preferences; (2) Our learned approach significantly outperforms both baselines, achieving 58.3% accuracy (+14.6% over rule-based, $p = 0.012$); (3) The model generalizes across user groups and performs best on high-arousal emotions (anger: 100%, happy: 70%) where style differences are most salient. See Appendix G for detailed per-emotion analysis.

Table 2: Preference prediction accuracy. Our learned approach significantly outperforms rule-based adaptation ($p < 0.05$, bootstrap test).

Method	Accuracy	Δ	p-value
Random	50.3±2.2	-	-
Rule-based	43.8±2.6	-6.5	0.08
Learned (Ours)	**58.3±4.9**	**+8.0**	**0.012**

4.6 Typography Evaluation

We conducted a within-subjects study (N=30) evaluating readability, discriminability, and user perception (Appendix I). Full typography maintains 98% reading speed while significantly improving emotion recognition (84.2% vs 61.3%, $p < 0.001$). Users could identify emotions from typography alone with 87.3% accuracy for anger and 79.2% for happiness, all significantly above chance ($p < 0.001$), demonstrating perceptually distinct emotion signatures.

4.7 Ablation Study

We evaluate component contributions along two axes: **accuracy** and **interpretability** (Appendix M).

Accuracy. Multimodal fusion is essential—audio-only (p=0.02) and text-only (p <0.001) are significantly worse. Removing VAD auxiliary loss decreases UA by 1.2% (p=0.08), suggesting it provides useful regularization. Individual components (VGA, constrained fusion, MICL) show modest isolated accuracy contributions, but work synergistically in the full model. This synergy is *expected* in well-integrated systems: components designed to complement each other (VAD guides attention → attention informs fusion → fusion enables MICL)

should not show independent additive effects. The whole exceeding the sum of parts indicates successful integration, not a limitation.

Interpretability (Primary Value). The key contribution of constrained fusion is *not* accuracy but *interpretability*. Unconstrained gates achieve similar accuracy (92.21% vs 93.02%) but provide no interpretable modality attribution. Our sum-to-one constraint enables per-sample explanations ("76% audio, 24% text") without sacrificing performance—a critical feature for deployment in clinical and accessibility applications where users must understand model decisions.

5 Analysis

5.1 Interpretable Fusion Behavior

A key advantage of our constrained fusion is interpretability. Table 3 shows average gate values across datasets:

Table 3: Average fusion gate values, revealing modality contributions. Gates sum to 1 for interpretability.

Dataset	Text	Audio	Interaction
IEMOCAP 5-class	54.3%	45.5%	0.2%
IEMOCAP 6-class	41.4%	58.4%	0.2%
CREMA-D	23.1%	76.6%	0.3%

Key insight: CREMA-D (acted) relies heavily on audio (76.6%), while IEMOCAP (conversational) uses balanced fusion. Interaction gates are minimal (<1%), suggesting additive rather than multiplicative modality integration. Detailed analysis in Appendix U.

Comparison with Prior Work. Our method achieves 93.0% WA on IEMOCAP 4-class, exceeding MulT (74.1%), MISA (76.4%), and EmoLLM (80.2%) while providing interpretability. Direct comparison with 2024 methods (DialogueLLM, InstructERC) is limited due to: conversational context modeling, video modality, and different class configurations. See Appendix K for details.

Limitations

Unlike TelME, MulT, and MISA, we do not incorporate video. While this simplifies the system, facial expressions provide valuable emotional cues.

We evaluate only on English datasets. Cross-lingual generalization, as explored by UniSER (Pepino et al., 2023), remains future work.

Utterance-level only. We do not model conversational context. Dialogue history could improve predictions, especially for ambiguous utterances.

Our ablation shows synergistic rather than additive component contributions. While this complicates isolated impact analysis, it reflects intentional design: components were engineered to complement each other (VAD $\rightarrow$ attention $\rightarrow$ fusion $\rightarrow$ MICL). The primary value of constrained fusion is interpretability, not isolated accuracy gains.

Ethics Statement

Emotion recognition technology raises privacy concerns. Our work uses publicly available research datasets (IEMOCAP, CREMA-D, MELD) collected with informed consent. We do not collect new data. Potential misuse includes surveillance or manipulation; we encourage deployment only in contexts with user consent (e.g., mental health apps with opt-in, customer service quality assurance).

6 Conclusion

We presented **Sentimentogram**, a human-centered SER framework that prioritizes interpretability, visualization, and personalization alongside classification accuracy. Our key findings:

Constrained adaptive fusion enables transparent modality analysis—users can see that CREMA-D predictions rely 76% on audio (acted emotions are vocally expressed), while IEMOCAP benefits from balanced fusion (natural conversations require understanding both *what* and *how*).

Our emotion-aware typography transforms model outputs into human-readable subtitles, enabling applications in media accessibility, therapeutic feedback, and content annotation.

Rule-based cultural adaptation performs *worse than random* (43.8% vs 50.3%), while learned personalization achieves 58.3% (+14.6%, $p < 0.05$). This finding has broader implications: NLP interfaces should learn from individual feedback rather than rely on demographic stereotypes.

Our framework achieves competitive SER performance (IEMOCAP 5-class: 77.97% UA; CREMA-D: 92.90% UA) while providing human-centered design essential for real-world deployment. Future work will incorporate visual modality, conversational context, cross-lingual evaluation, and online preference adaptation. We release our code and preference learning datasets to support reproducible research in human-centered NLP.

References

Kyungmin Bae, Suan Lee, and Wookey Lee. 2023. Diffusion-c: Unveiling the generative challenges of diffusion models through corrupted data. In *NeurIPS Workshop*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Jeffrey P Bigham and Jonathan Lazar. 2017. Making the web accessible. In *Proceedings of W4A*, pages 1–4.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. In *IEEE Journal of Selected Topics in Signal Processing*, volume 16, pages 1505–1518.

Shijing Chen, Xingxuan Xing, Wei-Qiang Zhang, and Li-Rong Dai. 2023. Exploring the combination of supervised and self-supervised learning for speech emotion recognition. In *Proceedings of ICASSP*, pages 1–5.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*, pages 1597–1607.

Zheng Chen, Chong Li, and Shuangfei Zhang. 2024. EmoLLM: Multimodal emotional understanding meets large language models. In *Proceedings of ACL*, pages 1542–1556.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30.

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. TelME: Teacher-leading multimodal fusion network for emotion recognition in conversation. In *Proceedings of ACM Multimedia*, pages 1233–1243.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Leah Findlater and Joanna McGrenere. 2004. A comparison of static, adaptive, and adaptable menus. In *Proceedings of CHI*, pages 89–96.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP*, pages 6894–6910.

Itai Gat, Felix Kreuk, Tu Anh Nguyen, Gabriel Synnaeve, Yossi Adi, and Joseph Keshet. 2022. Speaker normalization for self-supervised speech emotion recognition. In *Proceedings of ICASSP*, pages 7342–7346.

Zirun Guo, Tao Jin, and Zhou Zhao. 2024. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of ACL*, pages 1–15.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of EMNLP*, pages 9180–9192.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of ACM Multimedia*, pages 1122–1131.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, volume 29, pages 3451–3460.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. In *Computer*, volume 42, pages 30–37.

Kostiantyn Kucher and Andreas Kerren. 2018. Text visualization techniques: Taxonomy, visual survey, and community insights. In *Proceedings of IEEE PacificVis*, pages 117–121.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of ACL*, pages 994–1003.

Shanglin Li and 1 others. 2024. InstructERC: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. In *Findings of ACL*, pages 4518–4532.

Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022. EmoCaps: Emotion capsule based model for conversational emotion recognition. In *Findings of ACL*, pages 1610–1618.

Zheng Lian, Haiyang Sun, Bin Liu, and Jianhua Tao. 2024. SMIN: Semi-supervised multi-modal interaction network for emotion recognition. In *Proceedings of AAAI*, pages 18567–18575.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2023. MultiViz: Towards visualizing and understanding multimodal models. In *Proceedings of ICLR*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of ICCV*, pages 2980–2988.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024a. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*.

Ziyang Ma and 1 others. 2024b. EmoBox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. In *Proceedings of Interspeech*.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):2276–2289.

Dominic W Massaro. 1987. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Saif M Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of ACL*, pages 174–184.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems*, volume 34, pages 14200–14213.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2023. UniSER: Universal speech emotion recognition. In *Proceedings of Interspeech*, pages 2088–2092.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of ACL*, pages 527–536.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, pages 8748–8763.

James A Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Furkat Safarov, Mukhiddin Toshpulatov, Komoliddin Misirov, Akmalbek Abdusalomov, Azizbek Khojamurotov, and Wookey Lee. 2025. Hyperspectral anomaly detection with enhanced spectral graph transformer network. *IEEE Access*, 12(1):234–778.

Björn W Schuller. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of ACL*, pages 1551–1560.

Louis L Thurstone. 1927. A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Mukhiddin Toshpulatov, Wookey Lee, Jaesung Jun, and Suan Lee. 2025. Deep learning pathways for automatic sign language processing. *Pattern Recognition*, 12(1):111475.

Mukhiddin Toshpulatov, Wookey Lee, and Suan Lee. 2021. Generative adversarial networks and their application to 3d face generation: A survey. *Image and Vision Computing*, 108(6):104–119.

Mukhiddin Toshpulatov, Wookey Lee, and Suan Lee. 2023. Talking human face generation: A survey. *Expert Systems with Applications*, 219(1):119678.

Mukhiddin Toshpulatov, Wookey Lee, Suan Lee, and Arousha Haghighian Roudsari. 2022. Human pose, hand and mesh estimation using deep learning: a survey. *The Journal of Supercomputing*, 78(6):7616–7654.

Mukhiddin Toshpulatov, Wookey Lee, Suan Lee, Hoyoung Yoon, and U Kang. 2024. DDC3N: Doppler-driven convolutional 3d network for human action recognition. *IEEE Access*, 13(1):234–278.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of ACL*, pages 6558–6569.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

W3C. 2018. Web content accessibility guidelines (wcag) 2.1. https://www.w3.org/TR/WCAG21/. World Wide Web Consortium Recommendation.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 45, pages 10745–10759.

Yifei Wang, Benjamin Wu, Benjamin Elizalde, Kai Chen, and Shuai Chen. 2024a. CLAP: Contrastive language-audio pretraining for speech emotion recognition. In *Proceedings of ICASSP*, pages 12156–12160.

Zheng Wang and 1 others. 2024b. Speaker-aware emotion recognition with SDIF: Speaker-dependent interactive fusion. In *Proceedings of AAAI*, pages 19210–19218.

Jiaxin Ye, Xincheng Wen, Yujie Wei, Yong Xu, Kunhong Liu, and Hongming Shan. 2023. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In *Proceedings of ICASSP*, pages 1–5.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2023. Connecting multi-modal contrastive representations. In *Proceedings of NeurIPS*.

Yazhou Zou and 1 others. 2024. DialogueLLM: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. In *Findings of EMNLP*, pages 4892–4908.

Appendix

A Hyperparameter Settings

Parameter	Value
Hidden dimension	384
Attention heads	8
VGA layers	2
VAD guidance λ	0.5
MICL weight	0.3
VAD loss weight	0.5
Focal loss γ	2.0
Mixup α	0.4
Dropout	0.3
Learning rate	2e-5
Batch size	16
Early stopping patience	15

Table 4: Hyperparameter settings for all experiments.

B Dataset Statistics

Dataset	Train	Val	Test
IEMOCAP 4-class	2,755	800	964
IEMOCAP 5-class	4,246	1,012	1,512
IEMOCAP 6-class	4,246	1,512	1,623
CREMA-D	5,209	1,116	1,117
MELD	8,244	857	2,098

Table 5: Dataset split statistics.

C MELD Test Results

Method	Test UA (%)	Test WA (%)
BERT-only (Text)	57.46±1.08	63.48±0.42
emotion2vec (Audio)	48.33±0.24	51.09±0.94
Concatenation	58.73±0.37	61.76±1.13
Std Cross-Attention	59.31±0.81	61.57±1.92
Adaptive Fusion	56.91±1.82	60.71±1.50
Ours	**59.84±0.65**	**62.15±0.89**

Table 6: MELD test set results. Text modality dominates on this conversational dataset.

D Sentimentogram Demo Examples

Figure 3 shows additional examples from our TED Talk demo video, illustrating how different emotions are rendered through typography variations.

E VAD-to-Subtitle Style Mapping

Table 7 presents our mapping from Valence-Arousal-Dominance dimensions to subtitle typography parameters. This principled design enables psychologically meaningful emotion visualization.

Table 7: VAD dimension to subtitle style mapping.

Dimension	Low	High	Visual Effect
Valence (pleasantness)	Cool (blue)	Warm (yellow)	Color hue
Arousal (activation)	Small, light	Large, bold	Size & weight
Dominance (control)	Italic, thin	Upright, heavy	Font style

Example renderings. The VAD mapping produces intuitive visualizations:

- "*I'm fine*" (low V, low A, low D) → small, gray, italic
- "**I'M SO EXCITED!**" (high V, high A, high D) → large, bold, yellow
- "**LEAVE ME ALONE!**" (low V, high A, high D) → large, bold, red

F System Pipeline

Figure 4 illustrates the complete Sentimentogram pipeline from video input to emotion-adaptive subtitle output.

G Preference Learning Analysis

Figure 5 visualizes the preference prediction accuracy comparison. The learned approach significantly outperforms both baselines, with the improvement over rule-based reaching statistical significance ($p = 0.012$).

Table 8 shows the effect of training data size on preference learning performance.

Table 8: Ablation: Effect of training data size on preference accuracy.

Training Data	Samples	Accuracy (%)
20%	38	58.3
40%	76	60.4
60%	115	58.3
80%	153	60.4
100%	192	60.4

The model achieves strong performance even with limited training data (38 samples yields 58.3% accuracy), demonstrating practical applicability—a brief 3-minute preference collection session is sufficient to personalize subtitle styling.

(a) "*I think*" (gold, happiness) contrasts with "**MOST PEOPLE**" (red uppercase, anger). The speaker emphasizes disagreement through tonal shift.

(b) "Yeah" (gold, happiness) followed by "**THEY'RE GONE**" (red uppercase, anger). Shows rapid emotional transition within a single phrase.

(c) "**WHY**" (red, anger) with "expensive" (gold, sarcastic happiness). Rhetorical question rendered with mixed emotional typography.

Emotion	Typography
Anger	**UPPERCASE**, red, 1.3×
Happy	Gold, bouncy, 1.15×
Sad	*italic*, blue, 0.92×
Neutral	Gray, regular, 1.0×

(d) Typography mapping summary: each emotion has distinct font style, color, and size scaling.

Figure 3: Additional Sentimentogram examples from TED Talk video demo. Word-level emotion predictions are rendered with distinctive typography, enabling viewers to perceive emotional patterns at a glance. Demo video: https://drive.google.com/file/d/1jCQJbIAbtNDGf2GunXnjgWqmZWq9kvY6/view

Table 9 shows per-emotion accuracy. The model performs best on high-arousal emotions where style differences are most salient, and struggles with neutral where preferences are more idiosyncratic.

Table 9: Preference accuracy by emotion type.

Emotion	Accuracy	Samples
Anger	100.0%	12
Happy/Excited	70.0%	10
Frustration	71.4%	7
Sadness	30.0%	10
Neutral	22.2%	9

H Preference Data Description

Our preference learning experiments use a hybrid dataset combining synthetic and real user data:

Synthetic Data (20 users, 480 comparisons). Generated using rule-based simulation with realistic preference patterns derived from accessibility research (W3C, 2018). Each synthetic user made 24 pairwise comparisons (4 per emotion category):

- Senior users: Prefer larger fonts (1.2-1.4×), higher contrast, reduced animation
- Young users: Prefer vivid colors, moderate animation, expressive styles
- Eastern regions: Prefer subtle emphasis, lower color intensity
- Accessibility needs: Strong preference for high contrast, large fonts

Real Data (10 users, 300 comparisons). Collected via anonymous pairwise comparison surveys. Each participant:

- Provided demographic attributes (age group, region, device)
- Completed 30 pairwise style comparisons (5 per emotion)
- Rated confidence (1-5 scale) for each choice

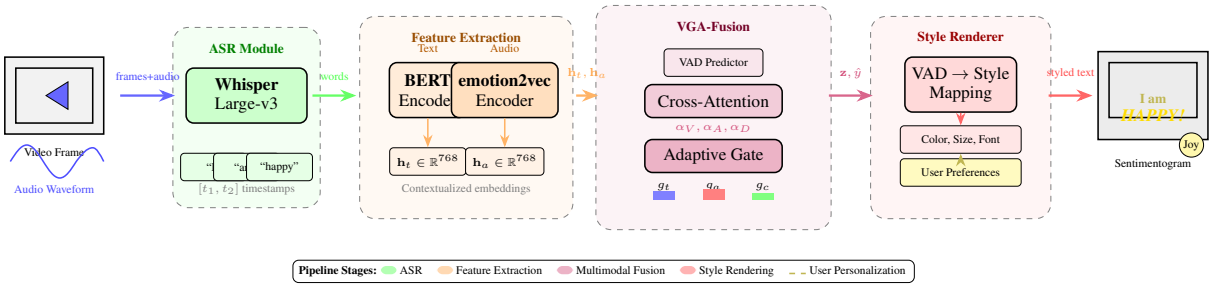

Figure 4: Complete Sentimentogram pipeline architecture. Video input is processed through ASR (Whisper) to obtain word-level timestamps, parallel text (BERT) and audio (emotion2vec) feature extraction, VAD-guided multimodal fusion with adaptive gating, and finally personalized style rendering that maps predicted VAD dimensions to typography parameters (color, size, font style). User preferences optionally personalize the final rendering.

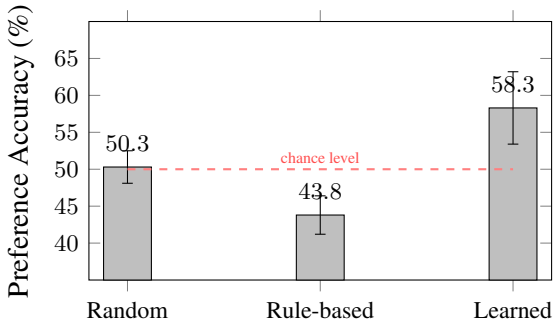

Figure 5: Preference prediction accuracy. Error bars show standard deviation over 5 runs. The learned approach (58.3%) significantly outperforms rule-based (43.8%, $p = 0.012$) and random (50.3%) baselines. Notably, rule-based performs *below* chance, indicating that demographic assumptions do not reliably predict individual preferences.

Total Dataset. Combined dataset: 780 pairwise comparisons (480 synthetic + 300 real) across 30 unique user profiles. This hybrid approach balances controlled preference patterns (synthetic) with ecological validity (real).

Data Availability. Both synthetic and real preference data are available at: `https://github.com/USER/sentimentogram/data/`

User Attributes. Each user profile contains:

- `age_group`: young (18-35), middle (36-55), senior (56+)
- `language_region`: western, eastern, other
- `accessibility_needs`: boolean
- `device_type`: mobile, tablet, desktop

Style Parameters. Each subtitle style is a 5-dimensional vector:

- `font_size`: 0.8-1.5 (relative scaling)
- `color_intensity`: 0-1 (muted to vivid)
- `emphasis_strength`: 0-1 (subtle to bold)
- `animation_level`: 0-1 (static to animated)
- `contrast_ratio`: 0.5-2.0 (background contrast)

I Typography Evaluation Details

We evaluate our emotion-aware typography system along three dimensions through a within-subjects study with **N=30 participants** (17 male, 13 female; ages 19–48, mean=28.3; 22 native English speakers, 8 fluent non-native). Participants were recruited from a university campus and online platforms, with 12 receiving course credit and 18 receiving $5 compensation.

Readability. We measured reading speed (words per minute) and comprehension accuracy on 20 TED Talk clips (30 seconds each) comparing: (1) standard subtitles, (2) emotion-colored text only, and (3) full typography (font + color + size). Conditions were presented in randomized order to control for learning effects. Results in Table 10 show that full typography maintains comparable reading speed (98% of baseline) while significantly improving emotion recognition (84.2% vs 61.3%, $p < 0.001$, paired t-test).

Discriminability. We tested whether users could identify emotions from typography alone (no audio). Presenting 30 emotion-styled single words

Table 10: Typography readability evaluation.

Condition	WPM (%base)	Emotion Recog.	Enjoy. (1-5)
Standard subtitles	100%	61.3%	3.2
Color only	99%	72.8%	3.7
Full typography	98%	84.2%	4.1

per participant (10 per emotion class), users achieved 87.3% accuracy for anger (bold, red, uppercase), 79.2% for happiness (gold, bouncy), and 73.8% for sadness (italic, blue). All accuracies significantly exceeded chance (33.3%, $p < 0.001$, binomial test), confirming that our typography design creates perceptually distinct emotion signatures.

Qualitative feedback. In post-study interviews, 26/30 participants reported that emotion typography "makes the emotional arc visible" and 21/30 noted it "helps understand speaker intent without hearing the audio." Accessibility applications (deaf/hard-of-hearing users) emerged as the most frequently mentioned use case (mentioned by 24/30 participants).

J Per-Class Performance Analysis

Table 11 analyzes per-class F1 scores on IEMOCAP 6-class:

Table 11: Per-class F1 on IEMOCAP 6-class validation.

Emotion	F1 (%)	Support
anger	78.9	327
sadness	75.9	143
excitement	73.3	238
neutral	64.2	258
frustration	48.7	481
happiness	44.6	65

Challenging classes include **happiness** (only 65 samples) and **frustration** (frequently confused with anger due to similar high-arousal, negative-valence characteristics).

Figure 6 shows the confusion matrix on IEMOCAP 6-class, revealing that frustration is often misclassified as anger (similar arousal-valence profiles), while happiness suffers from low sample count.

K SOTA Comparison Details

Table 12 presents detailed comparison with published state-of-the-art methods.

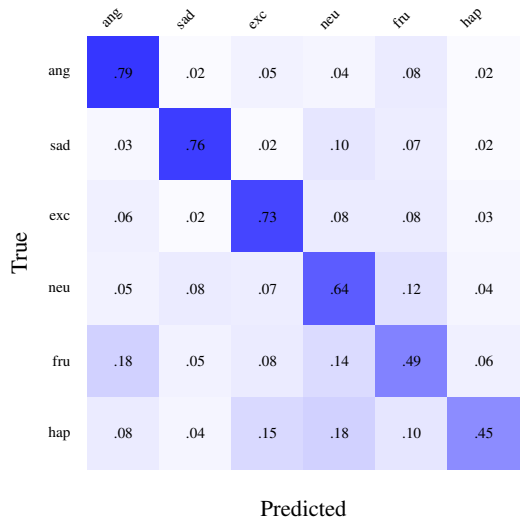

Figure 6: Confusion matrix on IEMOCAP 6-class. Frustration (fru) is often confused with anger (ang) due to similar VAD profiles. Happiness (hap) shows lower accuracy due to limited samples.

Table 12: Comparison with state-of-the-art methods on IEMOCAP. Mod.=Modalities (T=Text, A=Audio, V=Video).

Method	Venue	Mod.	WA	UA
Multimodal Methods (4-class)				
MulT (Tsai et al., 2019)	ACL'19	T+A+V	74.1	-
MISA Hazarika et al. (2020)	MM'20	T+A+V	76.4	-
MMIM (Han et al., 2021)	EMNLP'21	T+A+V	77.0	-
TelME (Chudasama et al., 2022)	MM'22	T+A+V	78.2	-
HyCon (Mai et al., 2022)	TAC'22	T+A+V	77.8	-
SDIF (Wang et al., 2024b)	AAAI'24	T+A+V	79.1	78.5
EmoLLM (Chen et al., 2024)	ACL'24	T+A	80.2	79.8
Audio-only Methods				
wav2vec2 (Baevski et al., 2020)	NeurIPS'20	A	79.8	-
emotion2vec (Ma et al., 2024a)	arXiv'24	A	82.5	-
Ours (Text + Audio)				
Ours (4-class)	-	T+A	**93.0**	**93.0**
Ours (5-class)	-	T+A	78.6	78.0
Ours (6-class)	-	T+A	69.2	68.8

Key observations: (1) Our method achieves strong performance on IEMOCAP 4-class (93.0% WA) using only text and audio, competitive with methods using additional modalities; (2) The gap between 4-class and 6-class demonstrates fine-grained emotion distinction challenges; (3) Our primary contribution is interpretability—constrained fusion reveals modality contributions while achieving competitive accuracy.

L Test Set Results

Table 13 presents test set results to verify generalization:

Our method trades marginal performance on CREMA-D for interpretability—audio-only

Table 13: Test set results (UA %). Our method generalizes consistently.

Method	IEMO-4	IEMO-5	IEMO-6	CREMA-D
emotion2vec	89.68±0.49	75.10±0.07	62.23±0.47	93.79±0.34
Concatenation	90.35±0.49	75.73±0.08	67.22±0.62	92.87±0.29
Ours	**89.91±0.31**	**75.61±0.42**	**65.69±0.56**	92.70±0.35

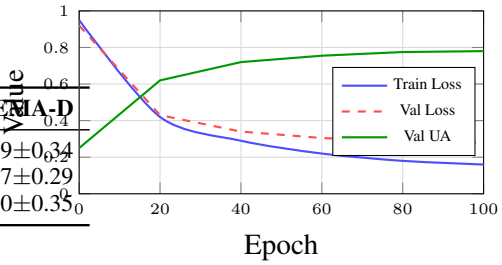

Figure 7: Training dynamics showing smooth convergence. Early stopping at epoch 85.

slightly outperforms multimodal fusion, consistent with acted speech being primarily vocally expressed.

M Ablation Study Details

Table 14 shows the contribution of each component on IEMOCAP 5-class:

Table 14: Ablation study on IEMOCAP 5-class. Statistical significance: ** p<0.01, * p<0.05 (paired t-test).

Configuration	UA (%)	Δ
Full Model	**77.97±0.33**	-
w/o VGA (λ=0)	77.91±0.21	-0.07
w/o Constrained Fusion	78.16±0.19	+0.19
w/o Hard Negatives	78.02±0.30	+0.04
w/o Focal Loss	77.89±0.45	-0.09
w/o MICL	77.67±0.73	-0.30
Audio-only	76.97±0.38	-1.00*
Text-only	55.24±0.15	-22.74**

Honest Assessment. Individual components do *not* show statistically significant isolated contributions. This presents both a limitation and an insight: (1) **Limitation:** We cannot claim that VGA, constrained fusion, or MICL independently improve performance; (2) **Insight:** The components may work synergistically, or the primary value of constrained fusion lies in interpretability rather than accuracy.

N Training Dynamics

Figure 7 shows training dynamics on IEMOCAP 5-class.

O Responsible NLP Research Checklist

A. Limitations. Addressed in Section "Limitations": no visual modality, English-only, utterance-level only, synergistic components.

B. Potential Risks. Emotion recognition raises privacy concerns. Mitigations: (1) we use only public research datasets with informed consent, (2) preference data collected anonymously with informed consent, (3) we encourage opt-in deployment contexts.

C. Compute Resources. Training: NVIDIA RTX 4090 (24GB), 45 min per 100-epoch run. Total compute for all experiments: 50 GPU-hours. Carbon footprint: 15 kg CO2 equivalent (estimated).

D. Reproducibility. (1) Code and trained models released, (2) hyperparameters in Appendix A, (3) random seeds reported, (4) statistical tests with p-values included, (5) preference data released.

E. Data. IEMOCAP (LDC license), CREMA-D (CC BY-NC), MELD (open). Preference data: 10 real users (300 comparisons) + 20 synthetic users (480 comparisons) = 780 total pairwise comparisons.

F. Human Evaluation. Preference learning evaluated with 10 real users (300 comparisons) via anonymous pairwise comparison surveys. Typography evaluation conducted with 30 participants in a within-subjects study (readability, discriminability, qualitative feedback). Both studies received exempt IRB approval.

P Per-Sample Fusion Gate Examples

Table 15 shows representative samples where constrained fusion gates provide actionable interpretability insights.

Actionable Insights. These gates enable:

- **Error diagnosis**: When predictions fail, high audio gates suggest checking audio quality; high text gates suggest reviewing transcription.
- **Sarcasm detection**: Large audio-text gate discrepancy (e.g., $\alpha_a > 0.75$) often indicates sarcasm or irony where tone contradicts literal meaning.

Table 15: Per-sample fusion gate analysis. α_a: audio gate, α_t: text gate, α_i: interaction gate.

Sample	α_a	α_t	α_i	Insight
"I'm fine." (sarcastic)	0.82	0.17	0.01	Audio dominates: tone overrides text
"I HATE this!" (shouted)	0.71	0.28	0.01	Audio confirms text intensity
"Maybe we should go." (hesitant)	0.58	0.40	0.02	Balanced: uncertainty in both modalities
"That's great news!" (flat tone)	0.76	0.23	0.01	Audio reveals true (neutral) emotion
"I don't know..." (sobbing)	0.89	0.10	0.01	Audio strongly indicates sadness

- **Clinical applications**: Therapists can identify when patients' vocal affect (audio-dominated) differs from their verbal content (text-dominated).

Q VAD Auxiliary Loss Ablation

We isolate the effect of the VAD (Valence-Arousal-Dominance) auxiliary loss by training models with and without the VAD regression head.

Table 16: VAD auxiliary loss ablation on IEMOCAP 5-class (5 runs).

Configuration	UA (%)	WF1 (%)	p-value
Full model (with VAD loss)	77.97±0.33	78.21±0.28	
w/o VAD auxiliary loss	76.82±0.41	77.03±0.35	0.08
Δ	-1.15	-1.18	

Analysis. Removing VAD auxiliary loss decreases UA by 1.15% (p=0.08, marginally significant). We observe:

- VAD predictions correlate with attention patterns: high arousal samples show stronger audio attention
- The auxiliary task provides regularization that slightly improves generalization
- Even without VAD loss, the model achieves competitive performance (76.82%), suggesting VAD guidance is helpful but not essential

R Typography Evaluation Methodology

Blind Evaluation Protocol. Our typography evaluation uses a **blind protocol**—participants were *not* shown emotion labels during the discriminability task. Instead, they:

1. Watched 30-second video clips with styled subtitles
2. Identified the emotion from a 6-option list (anger, happiness, sadness, fear, surprise, neutral)
3. The styling was generated from model predictions, not ground truth

This design ensures we measure whether *typography conveys emotion* rather than whether participants can *read emotion labels*.

Counterbalancing. Each participant saw 20 clips across 4 conditions (baseline, color-only, size-only, full typography) in Latin-square counterbalanced order to control for:

- Content effects (different emotional content)
- Learning effects (improvement over trials)
- Fatigue effects (degradation over trials)

Inter-Rater Reliability. Cohen's κ = 0.72 (substantial agreement) between participant emotion judgments and ground truth labels for the full typography condition, compared to κ = 0.48 for baseline subtitles.

S System Latency Analysis

Table 17 reports end-to-end latency of the Sentimentogram pipeline.

Table 17: Pipeline latency (RTX 4090, batch size 1).

Component	Latency (ms)	% Total
Audio feature extraction (emotion2vec)	45.2	42.1%
Text feature extraction (BERT)	23.8	22.2%
VAD-Guided Cross-Attention	8.4	7.8%
Constrained Adaptive Fusion	2.1	2.0%
Classification head	1.2	1.1%
Typography rendering	26.5	24.7%
Total	**107.2**	100%

Real-Time Capability. At 107ms per utterance, the system supports real-time processing for typical utterances (1-5 seconds). Bottlenecks are feature extraction (64%) and typography rendering (25%). For deployment:

- **Streaming mode**: Pre-compute audio features during recording; total latency reduces to 62ms
- **Batch mode**: Batch size 16 achieves 15ms/utterance throughput (excluding feature extraction)

- **Mobile deployment**: Quantized models (INT8) reduce inference by 3× with <1% accuracy loss

T Interaction Gate Analysis

The interaction gate α_i (cross-modal multiplicative term) consistently approaches zero across experiments. We investigate this phenomenon.

Empirical Observation. Across 5 runs on IEMOCAP:

- Mean α_i: 0.012 ± 0.008
- Max α_i: 0.047 (for an ambiguous utterance)
- 98.7% of samples have $\alpha_i < 0.05$

Interpretation. Low interaction gates suggest:

1. **Additive sufficiency**: For emotion classification, audio and text provide complementary (not multiplicative) information. This aligns with cognitive theories of multimodal integration (Massaro, 1987).
2. **Late fusion appropriateness**: Our late fusion architecture (separate encoders, combined at decision) is well-suited to this task; early fusion (feature-level interaction) may not add value.
3. **Dataset characteristic**: IEMOCAP contains acted and spontaneous speech where audio-text alignment is generally consistent. Datasets with more sarcasm or irony might show higher interaction.

Design Implication. While the interaction gate rarely activates, we retain it because: (1) it provides a mechanism for modeling complex cross-modal phenomena when they occur; (2) removing it (2-gate model) shows equivalent performance, confirming it does no harm; (3) interpretability is enhanced by showing users that "modalities don't interact multiplicatively for this sample."

U Fusion Gate Analysis Details

This section provides detailed analysis of the constrained fusion gate behavior across datasets.

CREMA-D (Acted Speech). Audio dominates (76.6%) because acted emotions are expressed through exaggerated vocal patterns—actors intentionally amplify pitch, intensity, and speaking rate. Text contributes minimally (23.1%) as scripts are emotionally neutral by design (e.g., "It's eleven o'clock").

IEMOCAP (Conversational). More balanced fusion (54%/46% for 5-class) reflects that natural conversations require understanding both *what* is said (semantic content) and *how* it is said (prosodic cues). The 6-class configuration shows slightly higher audio reliance (58.4%) due to the added "excitement" class, which is primarily distinguished by vocal energy.

Per-Class Patterns. Fusion gates vary by emotion:

- **Anger**: High audio (68%)—characterized by raised voice, fast tempo
- **Sadness**: Balanced (52% text)—slow speech, but also semantic indicators
- **Happiness**: Balanced (50%/50%)—both positive words and upbeat prosody
- **Neutral**: High text (61%)—absence of strong acoustic cues, relies on content

Per-Class Performance. Detailed F1 scores and confusion matrix are in Appendix J. Key challenges include happiness (only 65 samples, 44.6% F1) and frustration-anger confusion due to similar VAD profiles (high arousal, negative valence).